# MODEL SELECTION WITH GLMS

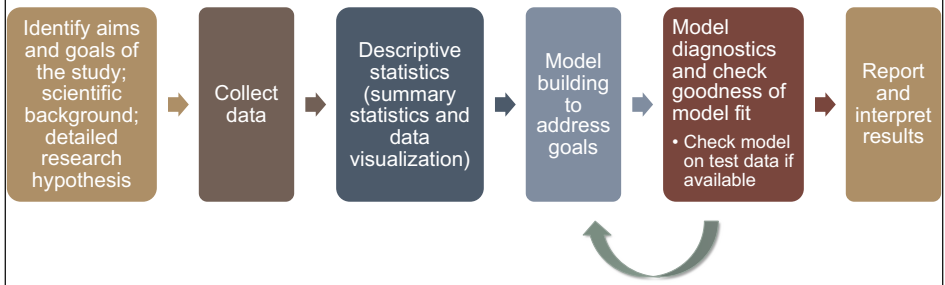
Chapter 5 Section 5.1-5.2

+ additional material



## Data Analysis Flow Chart

**IMPORTANT: THINK!!**






## What is your goal?

The goal of the study should guide your methods:

1. **Assessing** the relationship of a particular set of "focal" predictors with Y.
   - The focal predictors need to be in the model!
   - Include other predictors only to reduce probable sources of variability or control for known confounders.
   - But, including too many additional predictors could decrease precision.



## What is your goal?

**Assessing** the relationship of a particular set of "focal" predictors with Y.

Ex: Y = child IQ at age 3, X = average drinks (alcohol) per week of mother during pregnancy (focal predictor)

- May also want to control for: mother's education, parents' IQ, level of prenatal care, etc.
- But don't want too many variables in the model!
- Report relationship between X and Y "adjusting for…"

# What is your goal?

2. **Discovering** which predictors are related to Y.
   - Suppose the X's divide into a set of "active" predictors $X_A$ (good) and a set of "inactive" predictors $X_I$, (not helpful, but correlated with $X_A$).
   - Goal is to identify $X_A$, but not easy!
   - Ex: Is there a genetic component to obsesity? Y = BMI, X's are possible genes and lifestyle
3. **Predicting** future values of Y.
   - Need to be able to observe predictors for a new subject.



# Types of Studies

- Controlled experiments
- Controlled experiments with covariates
- Confirmatory observational studies
- Exploratory observational studies



# Controlled Experiments

- Investigators choose the values of the explanatory variable and randomly assign those values to experimental units.
- Explanatory variables specifically chosen since the investigator would like to know how they affect the response.
- The only model selection issue is whether or not to include interactions and the correct functional form.
- **Predictors generally should not be dropped from these studies**.



# Controlled Experiments with Covariates

- Investigators choose the values of the explanatory variable and randomly assign those values to experimental units.
- Other covariates (variables that may effect the response but are not randomly assigned) are measured.
- Explanatory variables specifically chosen since the investigator would like to know how they affect the response.
- **Again, the only model selection issue is whether or not to include interactions and the correct functional form.**
- **If covariates do not have a significant effect on the response, the investigator may choose to drop them from the model.**

## Confirmatory Observational Study

- Observational studies intended to test hypotheses derived from previous studies or from hunches.
- Data collected for explanatory variables and control variables specifically of interest to the study.
- Explanatory variables involved in hypotheses of interest sometimes called *primary variables*, and explanatory variables reflecting existing knowledge are *control variables* (or *known risk factors* in epidemiology).
- **Model selection for reduction of explanatory variables is not appropriate.**



## Exploratory Observational Study

- Observational study where investigators search for explanatory variables that might be related to the response variable.
- Could potentially have a very large set of potential predictor variables to include in the model.
- ***This is the case where we need good model selection tools!***



## Other Considerations

- More may not always be better.
- "Data snooping" – what works well for this data set may not work well in the population.
  - Ideally, we have a new set of "test data" to validate our model.
- Collinearity can distort interpretation of specific variables



## Other Considerations

- Your final model should *not* just be the result of automated model selection procedures.
- ***The model should be based on the science and previous research behind the problem.***
- The model will depend on the desired use of the model.
  - Descriptive use → typically emphasize precise estimation of regression coefficients.
  - Use for prediction → emphasize reducing prediction errors.

# What is the "best" model?

- "All models are wrong, but some are useful." – George Box
- There is never one "best" model – usually there is a subset of equivalently good models one can use.
- Principle of parsimony (Occam's razor): Simplest is best, if all else is equal!
- The identification of a subset of "good" predictors, the appropriate functional form of these predictors, and how to determine the final regression model constitute some of the most difficult problems in regression analysis.



# Key Questions in Model Building

- How do we know or choose which predictors to include in our model?
  - Is there a particular predictor of interest that needs to be in the model?
  - Are there known confounders or effect modifiers we need to include in the model?
- What form of the predictor should we use? Do we need a polynomial? transformation? interactions?
- How do we know if our model provides a good fit to the observed data?
- How do we know if our model makes good predictions of future data?



# Tools for answering these questions?

**Linear Models (Normal random component; Identity link):**

- Scatterplots and residual diagnostic plots for checking the form of the model and model assumptions
- t-tests and F-tests for comparing two nested models (reduced vs. full model)
- SSE or MSE = variability in the response leftover after accounting for the predictor variables
- $R^2$ = proportion of the variability in the response that can be explained by the model (equivalent to using SSE)
- Adjusted $R^2$ as a measure for comparing two models (equivalent to using MSE)



# Tools for answering these questions?

**Generalized Linear Model:**

- Scatterplots, side-by-side boxplots, residual plots
- Wald tests and Likelihood ratio tests (analysis of deviance) for comparing two nested models (reduced vs. full model)
- Residual deviance for goodness of fit if grouped data with fitted counts > 5
- Hosmer-Lemeshow goodness of fit test for binary data
- *Predictive models:*
  - Model selection criteria (e.g. AIC) and stepwise algorithms
  - ROC (receiver operating characteristic) curves